BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Multiple Regression

Overview

Multiple regression is used to account for (predict) the variance in an interval dependent, based on linear combinations of interval, dichotomous, or dummy independent variables. Multiple regression can establish that a set of independent variables explainsa proportion of the variance in a dependent variable at a significant level (significance test of R^2), and can establish the relative predictive importance of the independent variables (comparing beta weights). Power terms can be added as independent variables to explore curvilinear effects. Cross-product terms can be added as independent variables to explore interaction effects. One can test the significance of difference of two R^2's to determine if adding an independent variable to the model helps significantly. Using hierarchical regression, one can see how must variance in the dependent can be explained by one or a set of new independent variables, over and above that explained by an earlier set. Of course, the estimates (b coefficients and constant) can be used to construct a prediction equation and generate predicted scores on a variable for further analysis.

The multiple regression equation takes the form $y = b_1x_1 + b_2x_2 + ... + b_nx_n + c$. The b's are the regression coefficients, representing the amount the dependent variable y changes when the independent changes 1 unit. The c is the constant, where the regression line intercepts the y axis, representing the amount the dependent y will be when all the independent variables are 0. The standardized version of the b coefficients are the beta weights, and the ratio of the beta coefficients is the ratio of the relative predictive power of the independent variables. Associated with multiple regression is R^2, multiple correlation, which is the percent of variance in the dependent variable explained collectively by all of the independent variables.

Multiple regression shares all the assumptions of correlation: linearity of relationships, the same level of relationship throughout the range of the independent variable ("homoscedasticity"), interval or near-interval data, and data whose range is not truncated. In addition, it is important that the model being tested is correctly specified. The exclusion of important causal variables or the inclusion of extraneous variables can change markedly the beta weights and hence the interpretation of the importance of the independent variables.

Multiple regression with dummy variables yields the same inferences as multiple analysis of variance (MANOVA), to which it is statistically equivalent. When the dependent variable is a dichotomy the assumptions of multiple regression cannot be met and discriminant analysis or logistic regression is used instead. Partial least squares regression is sometimes used to predict one set of response variables from a set of independent variables.

Key Terms and Concepts

- The **regression equation** takes the form $Y = b_1*x_1 + b_2*x_2 + c + e$, where Y is the true dependent, the b's are the regression coefficients for the corresponding x (independent) terms, c is the constant or intercept, and e is the error term reflected in the residuals. Sometimes this is expressed more simply as $y = b_1*x_1 + b_2*x_2 + c$, where y is the estimated dependent and c is the constant (which includes the error term). Equations such as that above, with no interaction effects (see below), are called *main effects models.*
    - **Predicted values**, also called *fitted values*, are the values of each case based on using the regression equation for all cases in the analysis. In SPSS, dialog boxes use the term PRED to refer to predicted values and ZPRED to refer to standardized predicted values.

- **Adjusted predicted values** are the values of each case based on using the regression equation for all cases in the analysis except the given case.

- **Residuals** are the difference between the observed values and those predicted by the regression equation. In SPSS, dialog boxes use the term RESID to refer to residuals and ZRESID to refer to standardized residuals.

- **Dummy variables** are a way of adding the values of a nominal or ordinal variable to a regression equation. Each value of the categorical independent except one is entered as a dichotomy (ex., East = 1 if unit is in East, otherwise 0; West = 1 if unit is in West, otherwise 0; etc.). One class must be left out to prevent perfect multicollinearity in the model (see FAQ below).

- **Interaction effects** are sometimes called *moderator effects* because the interacting third variable which changes the relation between two original variables is a moderator variable which moderates the original relationship. For instance, the relation between income and conservatism may be moderated depending on the level of education.

  Interaction terms may be added to the model to incorporate the joint effect of two variables (ex., income and education) on a dependent variable (ex., conservatism) over and above their separate effects. One adds interaction terms to the model as crossproducts of the standardized independents and/or dummy independents, typically placing them after the simple "main effects" independent variables. In SPSS syntax mode, to create a new interaction variable X12 from variables X1 and X2, simply issue the command COMP X12 = X1*X2. In SAS, you can model directly: MODEL y = X1 + X2 +(X1*X2). Some computer programs will allow the researcher to specify the pairs of interacting variables and will do all the computation automatically. Crossproduct interaction terms may be highly correlated (multicollinear - see below) with the corresponding simple independent variables in the regression equation, creating problems with assessing the relative importance of main effects and interaction effects. *Note:* Because of possible multicollinearity, it may well be desirable to use *centered variables* (where one has subtracted the mean from each datum) -- a transformation which often reduces multicollinearity. Note also that there are alternatives to the crossproduct approach to analyzing interactions.

  The significance of an interaction effect is the same as for any other variable, except in the case of a set of dummy variables representing a single ordinal variable. When an ordinal variable has been entered as a set of dummy variables, the interaction of another variable with the ordinal variable will involve multiple interaction terms. In this case the F-test of the significance of the interaction of the two variables is the significance of the change of R-square of the equation with the interaction terms and the equation without the set of terms associated with the ordinal variable.

  An alternative common approach to interactions is to run separate regressions for each level of the interacting variable.

- **The regression coefficient, b,** is the average amount the dependent increases when the independent increases one unit and other independents are held constant. Put another way, the b coefficient is the slope of the regression line: the larger the b, the steeper the slope, the more the dependent changes for each unit change in the independent. The b coefficient is the unstandardized simple regression coefficient for the case of one independent. When there are two or more independents, the b coefficient is a *partial regression coefficient*, though it is common simply to call it a "regression coefficient" also.

  - **Interpreting b for dummy variables.** For b coefficients for dummy variables which have been *binary coded* (the usual 1=present, 0=not present method discussed above), b is relative to the *reference category* (the category left out). Thus for the set of dummy variables for "Region," assuming "North" is the reference category and education level is the dependent, a b of -1.5 for the dummy "South" means that the expected education level for the South is 1.5 years less than the average of "North" respondents. Dummy variables and their interpretation under alternative forms of coding is discussed below.

  - **Dynamic inference** is drawing the interpretation that the dependent changes b units because the independent

changes one unit. That is, one assumes that there is a change process (a dynamic) which directly relates unit changes in x to b changes in y. This assumption implies two further assumptions which may or may not be true: (1) b is stable for all subsamples or the population (*cross-unit invariance*) and thus is not an artificial average which is often unrepresentative of particular groups; and (2) b is stable across time when later re-samples of the population are taken (*cross-time invariance*).

- **t-tests** are used to assess the significance of individual b coefficients. specifically testing the null hypothesis that the regression coefficient is zero. A common rule of thumb is to drop from the equation all variables not significant at the .05 level or better. Note that *restricted variance* of the independent variable in the particular sample at hand can be a cause of a finding of no significance. Like all significance tests, the t-test assumes randomly sampled data. Note: t-tests are not used for dummy variables, even though SPSS and other statistical packages output them -- see Frequently Asked Questions section below. Note also that the t-test is a test only of the unique variance an independent variable accounts for, not of shared variance it may also explain, as shared variance while incorporated in R^2, is not reflected in the b coefficient.

  One- vs. two-tailed t tests. Also note that t-tests in SPSS and SAS are two-tailed, which means they test the hypothesis that the b coefficient is either significantly higher or lower than zero. If our model is such that we can rule out one direction (ex., negative coefficients) and thus should test only if the b coefficient is more than zero, we want a one-tailed test. The one-tailed significance level will be twice the two-tailed probability level: if SPSS reports .05, for instance, then the one-tailed equivalent significance level is .1.

- **Level-importance** is the b coefficient times the mean for the corresponding independent variable. The sum of the level importance contributions for all the independents, plus the constant, equals the mean of the dependent variable. Achen (1982: 72) notes that the b coefficient may be conceived as the "potential influence" of the independent on the dependent, while level importance may be conceived as the "actual influence." This contrast is based on the idea that the higher the b, the more y will change for each unit increase in b, but the lower the mean for the given independent, the fewer actual unit changes will be expected. By taking both the magnitude of b and the magnitude of the mean value into account, level importance is a better indicator of expected actual influence of the independent on the dependent.

- **The beta weights** are the regression (b) coefficients for standardized data. Beta is the average amount the dependent increases when the independent increases one standard deviation and other independent variables are held constant. If an independent variable has a beta weight of .5, this means that when other independents are held constant, the dependent variable will increase by half a standard deviation (.5 also). The ratio of the beta weights is the ratio of the estimated predictive importance of the independents. Note that the betas will change if variables or interaction terms are added or deleted from the equation. Reordering the variables without adding or deleting will not affect the beta weights. That is, the beta weights help assess the relative importance of the independent variables relative to the given model embodied in the regression equation.

  Note that the betas reflect the unique contribution of each independent variable. Joint contributions contribute to R-square but are not attributed to any particular independent variable. The result is that the betas may underestimate the importance of a variable which makes strong joint contributions to explaining the dependent variable but which does not make a strong unique contribution. Thus when reporting relative betas, one must also report the correlation of the independent variable with the dependent variable as well, to acknowledge if it has a strong correlation with the dependent variable.

  - **Standardized** means that for each datum the mean is subtracted and the result divided by the standard deviation. The result is that all variables have a mean of 0 and a standard deviation of 1. This enables comparison of variables of differing magnitudes and dispersions. Only standardized b-coefficients (beta weights) can be compared to judge relative predictive power of independent variables.

  - Note some authors use "b" to refer to sample regression coefficients, and "beta" to refer to regression coefficients for population data. They then refer to "standardized beta" for what is simply called the "beta

weight" here.

- When assessing the relative importance of independents, light is thrown on the ratio of beta weights by also looking at the correlation and semipartial (part) correlations of a given independent with the dependent.
  - Correlation: Pearson's r^2 is the percent of variance in the dependent explained by the given independent when (unlike the beta weights) all other independents are allowed to vary. The result is that the magnitude of r^2 reflects not only the unique covariance it shares with the dependent, but uncontrolled effects on the dependent attributable to covariance the given independent shares with other independents in the model.
  - Semipartial correlation, also called part correlation: Semipartial correlation, reported as "part corr" by SPSS, in its squared form is the percent of variance in the dependent uniquely attributable to the given independent when other variables in the equation are controlled (not allowed to vary).

- **The intercept**, variously expressed as e, c, or x-sub-0, is the estimated Y value when all the independents have a value of 0. Sometimes this has real meaning and sometimes it doesn' t — that is, sometimes the regression line cannot be extended beyond the range of observations, either back toward the Y axis or forward toward infinity.

- **R^2**, also called *multiple correlation* or the *coefficient of multiple determination*, is the percent of the variance in the dependent explained uniquely or jointly by the independents. R-squared can also be interpreted as the proportionate reduction in error in estimating the dependent when knowing the independents. That is, R^2 reflects the number of errors made when using the regression model to guess the value of the dependent, in ratio to the total errors made when using only the dependent's mean as the basis for estimating all cases. Mathematically, R^2 = (1 - (SSE/SST)), where SSE = error sum of squares = SUM((Yi - EstYi)squared), where Yi is the actual value of Y for the ith case and EstYi is the regression prediction for the ith case; and where SST = total sum of squares = SUM((Yi - MeanY)squared). In summary, R^2 is 1 minus regression error as a percent of total error and will be 0 when regression error is as large as it would be if you simply guessed the mean for all cases of Y.
  - **Squared semipartial (part) correlation**: the proportion of total variance in a dependent variable explained uniquely by a given independent variable, not counting joint explanation. When the given independent variable is removed from the equation, R^2 will be reduced by this amount. Likewise, it may be interpreted as the amount R^2 will increase when that independent is added to the equation. R^2 minus the sum of all squared semi-partial correlations is the variance explained jointly by all the independents (the "shared variance"of the model).
  - **Squared partial correlation**: the proportion of variance in the dependent not explained by other independents, but which is explained uniquely by the given independent variable (that is, not counting jointly explained variance).
  - **Maximizing R^2 by adding variables** is inappropriate unless variables are added to the equation for sound theoretical reason. At an extreme, when n-1 variables are added to a regression equation, R^2 will be 1, but this result is meaningless. Adjusted R^2 is used as a conservative reduction to R^2 to penalize for adding variables and is required when the number of independent variables is high relative to the number of cases..
  - **R^2 and differences in variances between samples**. R^2, like other forms of correlation, is sensitive to restricted variance. Achen (1982: 75) gives the example of a study of the influence of a measure of a newspaper's bias in stories,. Bias was used to predict votes for candidates in primary and general elections. The correlation for general elections was .84 and for primary elections was .64, tempting the wrong conclusion that newspaper bias was more influential in general elections. However, the variance of bias was much less in the primaries than in the general elections. The greater variance in general elections allowed

more explanation of the variance in votes in general elections, especially since the general elections had less variance to explain. However, the primaries exhibited a higher b coefficient (that is, an additional biased story in the primaries had a greater impact on the percent of votes for the candidate). Achen thus warns that R^2 's cannot be compared between samples due to differences in variances of the independent and dependent variables.

- **Adjusted R-Square** is an adjustment for the fact that when one has a large number of independents, it is possible that R^2 will become artificially high simply because some independents' chance variations "explain" small parts of the variance of the dependent. At the extreme, when there are as many independents as cases in the sample, R^2 will always be 1.0. The adjustment to the formula arbitrarily lowers R^2 as p, the number of independents, increases. Some authors conceive of adjusted R^2 as the percent of variance "explained in a replication, after subtracting out the contribution of chance." When used for the case of a few independents, R^2 and adjusted R^2 will be close. When there are a great many independents, adjusted R^2 may be noticeably lower. The greater the number of independents, the more the researcher is expected to report the adjusted coefficient.

  Adjusted $R^2 = 1 - ((1-R^2)(N-1 / N - k - 1))$.

  where n is sample size and k is the number of terms in the model not counting the constant (i.e., the number of independents).

- **R^2 increments** usually refer to the amount R^2 decreases when a variable is dropped from the equation. Indeed, the *R^2 difference test* refers to running regression for a full model and for the model minus one variable, then subtracting the R>sup>2's and testing the significance of the difference. Since stepwise regression adds one variable at a time to the regression model, generating an R^2 value each time, subtracting each R^2 from the prior one also gives the R^2 increment. R^2 increments are tested by the F-test and are intrinsic to hierarchical regression, discussed below.

  $F\text{-incremental} = [(R^2_{with} - R^2_{without})/m] / [(1 - R^2)/df]$

  where m = number of IVs in new block which is added; and df = N - k - 1 (where N is sample size; k is number of indpendent variables). F is read with m and df degrees of freedom to obtain a p (probability) value. Note the without model is nested within the with model.

  - **Relation of beta weights to R^2 and R^2 increments.** Some authors state that the ratio of the squared beta weights indicates each independent variable's R-square increment. This is incorrect.

    The beta weights for the equation in the final step of stepwise regression do not partition R^2 into increments associated with each independent because beta weights are affected by which variables are in the equation. The beta weights estimate the relative predictive power of each independent, controlling for all other independent variables in the equation for a given model. The R^2 increments estimate the predictive power an independent brings to the analysis when it is added to the regression model, as compared to a model without that variable. Beta weights compare independents in one model, whereas R^2 increments compare independents in two or more models.

    This means that assessing a variable's importance using R^2 increments is very different from assessing its importance using beta weights. The magnitude of a variable's beta weight reflects its relative explanatory importance controlling for other independents in the equation. The magnitude of a variable's R^2 increment reflects its additional explanatory importance given that common variance it shares with other independents entered in earlier steps has been absorbed by these variables. For causal assessments, beta weights are better (though see the discussion of corresponding regressions for causal analysis). For purposes of sheer prediction, R^2 increments are better.

- **Standard Error of Estimate (SEE), confidence intervals, and prediction intervals**. Confidence intervals around the mean are discussed in the section on significance. In regression, however, the confidence refers to more than one thing. Note the confidence and prediction intervals will improve (narrow) if sample size is increased, or the confidence level is decreased (ex., from 95% to 90%).

  - **The confidence interval of the regression coefficient**. Based on t-tests, the confidence interval is the plus/minus range around the observed sample regression coefficient, within which we can be, say, 95% confident the real regression coefficient for the population regression lies. Confidence limits are relevant only to random sample datasets. If the confidence interval includes 0, then there is no significant linear relationship between x and y. We then do not reject the null bypothesis that x is independent of y. In SPSS 10, select Analyze - Regression - Linear - Statistics - check "Confidence intervals"

  - **The confidence interval of y** (the dependent variable), also called the standard error of mean prediction. For the 95% confidence limits, the confidence interval on y is plus/minus 1.96 times the standard error of estimate (SEE). SEE is SQRT(RSS/df), where RSS is the sum of squares of residuals and df is degrees of freedom = (n - m -1), where n is sample size and m is the number of independent variables. Some 95 times out of a hundred, the true mean of y will be within the confidence limits around the observed mean of n sampled cases. Note the confidence interval of y deals with the mean, not an individual case of y. Moreover, the confidence interval is narrower than the prediction interval, which deals with individual cases. Note a number of textbooks do not distinguish between confidence and prediction intervals and confound this difference. In SPSS 10, select Analyze - Regression - Linear - Statistics - Save - and under "Prediction intervals" check "Mean" - under "Confidence interval" set the confidence level you want (ex., 95%). Note SPSS calls this a *prediction interval for the mean.*

  - **The prediction interval of y**. For the 95% confidence limits, the prediction interval on a fitted value is plus/minus is the estimated value plus or minus 1.96 times SQRT(SEE + S^2_y), where S^2_y is the standard error of the mean prediction. Thus some 95 times out of a hundred, a case with the given values on the independent variables would lie within the computed prediction limits. The prediction interval will be wider (less certain) than the confidence interval, since it deals with an interval estimate of cases, not means. In SPSS 10, select Analyze - Regression - Linear - Statistics - Save - and under "Prediction intervals" check "Individual" - under "Confidence interval" set the confidence level you want (ex., 95%).

- **F test**: The F test is used to test the significance of R, which is the same as testing the significance of R^2, which is the same as testing the significance of the regression model as a whole. If prob(F) < .05, then the model is considered significantly better than would be expected by chance and we reject the null hypothesis of no linear relationship of y to the independents. F is a function of R^2, the number of independents, and the number of cases. F is computed with k and (n - k - 1) degrees of freedom, where k = number of terms in the equation not counting the constant.
  $F = [R^2/k]/[(1 - R^2)/(n - k - 1)]$.

  In SPSS, the F test appears in the ANOVA table, which is part of regression output. Note that the F test is too lenient for the stepwise method of estimating regression coefficients and an adjustment to F is recommended (see Tabachnick and Fidell, 2001: 143 and Table C.5).

- **Partial F test**: Partial-F can be used to assess the significance of the difference of two R^2's for nested models. Nested means one is a subset of the other, as a model with interaction terms and one without. Also, unique effects of individual independents can be assessed by running a model with and without a given independent, then taking partial F to test the difference. In this way, partial F plays a critical role in the trial-and-error process of model-building.

  Let there be q be a larger model and let p be a nested smaller model.
  Let RSSp be the residual sum of squares (deviance) for the smaller model.

Let RSSq be the residual sum of squares for the larger model.
Partial F has df1 and df2 degress of freedom, where
df1 = df for RSSp minus RSSq
df2 = df of RSS in the larger model
Partial F = (RSSp - RSSq)/(df1*[RSSq/df2])

In SPSS 10, run Analyze - Regression - Linear for the larger and smaller models. The ANOVA table, printed as default output, will list RSS and the corresponding df. Plug these values into the equation above to computer partial F for testing the difference between the models, then find the prob(F) in an F table with df1, df2 degrees of freedom. An alternate equation for testing the difference of models is given below.

- **OLS** stands for ordinary least squares. This derives its name from the criterion used to draw the best fit regression line: a line such that the sum of the squared deviations of the distances of all the points to the line is minimized.

- **Outliers** are data points which lie outside the general linear pattern of which the midline is the regression line. A rule of thumb is that outliers are points whose standardized residual is greater than 3.3 (corresponding to the .001 alpha level). SPSS will list these if you ask for "casewise diagnostics" under the Statistics button. The removal of outliers from the data set under analysis can at times dramatically affect the performance of a regression model. Outliers should be removed if there is reason to believe that other variables not in the model explain why the outlier cases are unusual -- that is, these cases need a separate model. Alternatively, outliers may suggest that additional explanatory variables need to be brought into the model (that is, the model needs respecification). Another alternative is to use *robust regression*, whose algorithm gives less weight to outliers but does not discard them.

  - The **leverage statistic, h**, also called the *hat-value*, is available to identify cases which influence the regression model more than others. The leverage statistic varies from 0 (no influence on the model) to 1 (completely determines the model). A rule of thumb is that cases with leverage under .2 are not a problem, but if a case has leverage over .5, the case has undue leverage and should be examined for the possibility of measurement error or the need to model such cases separately.

  - **Cook's distance, D**, is another measure of the influence of a case (see the output example). Cases with larger D values than the rest of the data are those which have unusual leverage.Fox (1991: 34) suggests as a cut-off for detecting influential cases, values of D greater than 4/(n - k - 1), where n is the number of cases and k is the number of independents.

  - **Studentized residuals** are also used to detect outliers with high leverage. The studentized residual is also called the *deleted studentized residual* because its calculation involves leaving out one case in turn for each of the cases. Other terms include *externally studentized residual* or, misleadingly, *standardized residual*. In a plot of studentized residuals, one may draw lines at plus and minus two standard units to highlight cases outside the range where 95% of the cases normally lie.

  - **Partial regression plots**, also called *partial regression leverage plots* or *added variable plots*, are yet another way of detecting influential sets of cases. Partial regression plots are a series of bivariate regression plots of the dependent variable with each of the independent variables in turn. The plots show cases by number or label instead of dots. One looks for cases which are outliers on all or many of the plots.

- **Multicollinearity** is the intercorrelation of independent variables. R^2's near 1 violate the assumption of no perfect collinearity, while high R^2's increase the standard error of the beta coefficients and make assessment of the unique role of each independent difficult or impossible. While simple correlations tell something about multicollinearity, the preferred method of assessing multicollinearity is to regress each independent on all the other independent variables in the equation. Inspection of the correlation matrix reveals only bivariate multicollinearity, for bivariate correlations > .90. To assess multivariate multicollinearity, one uses tolerance or VIF, which build in the regressing of each independent on all the others.. Even when multicollinearity is present, note that estimates for

other variables in the equation (variables which are not collinear with others) are not affected.

Note that a corollary is that very high standard errors of b coefficients is an indicator of multicollinearity in the data

- **Tolerance** is is 1 - R^2 for the regression of that independent variable on all the other independents, ignoring the dependent. There will be as many tolerance coefficients as there are independents. The higher the intercorrelation of the independents, the more the tolerance will approach zero. As a rule of thumb, if tolerance is less than .20, a problem with multicollinearity is indicated.

  When tolerance is close to 0 there is high multicollinearity of that variable with other independents and the b and beta coefficients will be unstable.The more the multicollinearity, the lower the tolerance, the more the standard error of the regression coefficients. Tolerance is part of the denominator in the formula for calculating the confidence limits on the b (partial regression) coefficient.

- **Variance-inflation factor, VIF** *VIF* is the variance inflation factor, which is simply the reciprocal of tolerance. Therefore, when VIF is high there is high multicollinearity and instability of the b and beta coefficients. VIF and tolerance are found in the SPSS output section on *collinearity statistics*. The table below shows the inflationary impact on the standard error of the regression coefficient (b) of the jth independent variable for various levels of multiple correlation (R_j), tolerance, and VIF (adapted from Fox, 1991: 12). Note that in the "Impact on SE" column, 1.0 corresponds to no impact, 2.0 to doubling the standard error, etc.:

R_j	Tolerance	VIF	Impact on SE_b
0	1	1	1.0
.4	.84	1.19	1.09
.6	.64	1.56	1.25
.75	.44	2.25	1.5
.8	.36	2.78	1.67
.87	.25	4.0	2.0
.9	.19	5.26	2.29

  Standard error is doubled when VIF is 4.0 and tolerance is .25, corresponding to R_j = .87. Therefore VIF >= 4 is an arbitrary but common cut-off criterion for deciding when a given independent variable displays "too much" multicollinearity: values above 4 suggest a multicollinearity problem. Some researchers use the more lenient cutoff of 5.0: if VIF >=5, then multicollinearity is a problem.

  In SPSS 10, to compute VIF and tolerance values, go to Analyze - Regression - Linear - Dependent (input dependent variable) - Independent (s) (input explanatory variables) - Statistics - check "Collinearity diagnostics" - Continue - OK. Tolerance and VIF appear in the "Coefficients" table.

- The **collinearity diagnostics table** in SPSS is an alternative method of assessing if there is too much multicollinearity in the model. To simplify, crossproducts of the independent variables are factored. High eigenvalues indicate dimensions (factors) which account for a lot of the variance in the crossproduct matrix. Eigenvalues close to 0 indicate dimensions which explain little variance. Multiple eigenvalues close to 0 indicate an *ill-conditioned crossproduct matrix*, meaning there may be a problem with multicollinearity and the condition indices should be examined as described below.

  - **Condition indices**. Condition indices are used to flag excessive collinearity in the data. A condition

index over 30 suggests serious collinearity problems and an index over 15 indicates possible collinearity problems. If a factor (component) has a high condition index, one looks in the *variance proportions* column to see if that factor accounts for a sizable proportion of variance in two or more variables (that is, if two or more variables are most heavily loaded on that factor). If this is the case, these variables have high linear dependence and multicollinearity is a problem, with the effect that small data changes or arithmetic errors may translate into very large changes or errors in the regression analysis. Note that it is possible for the rule of thumb for condition indices (no index over 30) may indicate multicollinearity, even when the rules of thumb for tolerance or VIF =4) are met. Computationally, a "singular value" is the square root of an eigenvalue, and "condition indices" are the ratio of the largest singular values to each other singular value.

- **Stepwise multiple regression**, also called *statistical regression*, is a way of computing OLS regression in stages. In stage one, the independent best correlated with the dependent is included in the equation. In the second stage, the remaining independent with the highest partial correlation with the dependent, controlling for the first independent, is entered. This process is repeated, at each stage partialling for previously-entered independents, until the addition of a remaining independent does not increase R-squared by a significant amount (or until all variables are entered, of course). Alternatively, the process can work backward, starting with all variables and eliminating independents one at a time until the elimination of one makes a significant difference in R-squared.

  Note that if one is using sets of dummy variables, the stepwise procedure must be performed manually as SPSS (at least through version 10) has no way to add/remove blocks of variables automatically but rather will treat each dummy as if it were an ordinary variable. That is, if using dummy variables one must run a series of manually-created equations which add/remove sets of dummy variables as a block.

  Stepwise regression is used in the exploratory phase of research or for purposes of pure prediction,, not theory testing. In the theory testing stage the researcher should base selection of the variables and their order on theory, not on a computer algorithm. Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained." Likewise, the nominal .05 significance level used at each step in stepwise regression is subject to inflation, such that the real significance level by the last step may be much worse, even below .50, dramatically increasing the chances of Type I errors. See Draper, N.R., Guttman, I. & Lapczak, L. (1979). For this reason, Fox (1991: 18) strongly recommends any stepwise model be subjected to cross-validation.

- **Hierarchical multiple regression** is similar to stepwise regression, but the researcher, not the computer, determines the order of entry of the variables. F-tests are used to compute the significance of each added variable (or set of variables) to the explanation reflected in R-square. This hierarchical procedure is an alternative to comparing betas for purposes of assessing the importance of the independents. In more complex forms of hierarchical regression, the model may involve a series of intermediate variables which are dependents with respect to some other independents, but are themselves independents with respect to the ultimate dependent. Hierarchical multiple regression may then involve a series of regressions for each intermediate as well as for the ultimate dependent.

  For hierarchical multiple regression, in SPSS first specify the dependent variable; then enter the first independent variable or set of variables in the independent variables box; click on "Next" to clear the IV box and enter a second variable or set of variables; etc. One also clicks on the Statistics button and selects "R-squared change." Note that the error term will change for each block or step in the hierarchical analysis. If this is not desired, it can be avoided by selecting Statistics, General Linear Model, GLM-General Factorial, then specifying Type I sums of squares. This will yield GLM results analogous to hierarchical regression but with the same error term across blocks.

Assumptions

- **Proper specification of the model**: If relevant variables are omitted from the model, the common variance they share with included variables may be wrongly attributed to those variables, and the error term is inflated. If causally irrelevant variables are included in the model, the common variance they share with included variables may be wrongly attributed to the irrelevant variables. The more the correlation of the irrelevant variable(s) with other independents, the greater the standard errors of the regression coefficients for these independents. Omission and irrelevancy can both affect substantially the size of the b and beta coefficients. This is one reason why it is better to use regression to compare the relative fit of two models rather than to seek to establish the validity of a single model specification.

  Sometimes this is phrased as the assumption that "independent variables are measured without error." Error attributable to omitting causally important variables means that, to the extent that these unmeasured variables are correlated with the measured variables which are in the model, the b coefficients will be off. If the correlation is positive, then b coefficients will be too high; if negative, too low. That is, when a causally important variables is added to the model, the b coefficients will all change, assuming that variable is correlated with existing measured variables in the model (usually the case).

  The specification problem in regression is analogous to the problem of spuriousness in correlation, where a given bivariate correlation may be inflated because one has not yet introduced control variables into the model by way of partial correlation.

  Suppression. Note that when the omitted variable has a suppressing effect, coefficients in the model may underestimate rather than overestimate the effect of those variables on the dependent. Suppression occurs when the omitted variable has a positive causal influence on the included independent and a negative influence on the included dependent (or vice versa), thereby masking the impact the independent would have on the dependent if the third variable did not exist.

- **No overfitting**. The researcher adds variables to the equation while hoping that adding each significantly increases R-squared. However, there is a temptation to add too many variables just to increase R-squared by trivial amounts. Such overfitting trains the model to fit noise in the data rather than true underlying relationships. Subsequent application of the model to other data may well see substantial drops in R-squared.

  - **Cross-validation** is a strategy to avoid overfitting. Under cross-validation, a sample (typically 60% to 80%) is taken for purposes of training the model, then the hold-out sample (the other 20% to 40%) is used to test the stability of R-squared. This may be done iteratively for each alternative model until stable results are achieved.

- **Continuous data** are required (interval or ratio), though it is common to use ordinal data. Dummy variables form a special case and are allowed in OLS regression as independents. Dichotomies may be used as independents but not as the dependent variable. Use of a dichotomous dependent in OLS regression violates the assumptions of normality and homoscedasticity as a normal distribution is impossible with only two values. Also, when the values can only be 0 or 1, residuals will be low for the portions of the regression line near Y=0 and Y=1, but high in the middle -- hence the error term will violate the assumption of homoscedasticity (equal variances) when a dichotomy is used as a dependent. Even with large samples, standard errors and significance tests will be in error because of lack of homoscedasticity.

- **Unbounded data** are an assumption. That is, the regression line produced by OLS can be extrapolated in both directions but is meaningful only within the upper and lower natural bounds of the dependent.

- **Data are not censored, sample selected, or truncated.** There are as many observations of the independents as for the dependents. Click here for further discussion of this assumption.

- **Absence of perfect multicollinearity**. When there is perfect multicollinearity, there is no unique regression solution. Perfect multicollinearity occurs if independents are linear functions of each other (ex., age and year of

birth), when the researcher creates dummy variables for all values of a categorical variable rather than leaving one out, and when there are fewer observations than variables.

- **Absence of high multicollinearity**. When there is high but imperfect multicollinearity, a solution is still possible but as the independents increase in correlation with each other, the standard errors of the regression coefficients will become inflated. High multicollinearity does not change the estimates of the coefficients, only their reliability. This means that it becomes difficult to assess the relative importance of the independent variables using beta weights.

- **Linearity**. Regression analysis is a linear procedure. To the extent nonlinear relationships are present, conventional regression analysis will underestimate the relationship. That is, R-square will underestimate the variance explained overall and the betas will underestimate the importance of the variables involved in the non-linear relationship.

  When nonlinearity is present, there may be a need for exponential or interactive terms. Nonlinear transformation of selected variables may be a pre-processing step, but this is not common because it runs the danger of overfitting the model to what are, in fact, chance variations in the data. Power and other transform terms should be added only if there is a theoretical reason to do so. Adding such terms runs the risk of introducing multicollinearity in the model. A guard against this is to use centering when introducing power terms (subtract the mean from each score). Correlation and unstandardized b coefficients will not change as the result of centering.

  In regression, as a rule of thumb, nonlinearity is generally not a problem when the standard deviation of the dependent is more than the standard deviation of the residuals. Linearity is further discussed in the section on data assumptions. Note also that regression smoothing techniques and nonparametric regression exist to fit smoothed curves in a nonlinear manner.

  - The **same underlying distribution** is assumed for all variables. To the extent that an independent variable has a different underlying distribution compared to the dependent (bimodal vs. normal, for instance), then a unit increase in the independent will have nonlinear impacts on the dependent. Even when independent/dependent data pairs are ordered perfectly, unit increases in the independent cannot be associated with fixed linear changes in the dependent. For instance, perfect ordering of a bimodal independent with a normal dependent will generate an s-shaped scatterplot not amenable to a linear solution. Linear regression will underestimate the correlation of the independent and dependent when they come from different underlying distributions.

    Transforms are sometimes used to force all variables to a normal distribution. For instance, square root, logarithmic, and inverse (x = 1/x) transforms correct a positive skew distribution, while powers correct a negative skew distribution. Such transforms lose information about the original scale of the variable, however.

  - **Partial regression plots** are often used to assess nonlinearity. These are simply plots of each independent on the x axis against the dependent on the y axis. Curvature in the pattern of points in a partial regression plot shows if there is a nonlinear relationship between the dependent and any one of the independents taken individually. Note, however, that whereas partial regression plots are preferred for illuminating cases with high leverage, partial residual plots (below) are preferred for illuminating nonlinearities.

  - **Partial residual plots**, also called *component-plus-residual plots*, are a preferred method of assessing nonlinearity. A *partial residual* for a given independent is the residual plus the product of the b coefficient times the value of a case on that independent. That is, partial residuals add the linear component of an independent back into the residual (hence component-plus-residual plots). The partial residual plot shows a given independent on the y axis and the corresponding partial residual on the x axis. There is one partial residual plot per independent variable. Partial residual plots have the advantage over partial regression plots in that the y axis incorporates a b coefficient, which in turn reflects both the independent variable and

control effects on that variable by other independents. The slope of the partial residuals will be the same as for the regression, but a **lowess smoothing** line may be drawn to highlight curvature of the data. The curvature of partial regression plots can illustrate both monotone (falling or rising) and nonmonotone (falling and rising) nonlinearities.

- **Simple residual plots** also show nonlinearity but do not distinguish monotone from nonmonotone nonlinearity. These are plots of standardized residuals against standardized estimates of Y, the dependent variable. In SPSS this is ZRESID vs. ZPRED. The plot should show a random pattern, with no nonlinearity or heteroscedasticity. In jargon, this will show the error vector is orthogonal to the estimate vector. Non-linearity is, of course, shown when points form a curve. Non-normality is shown when points are not equally above and below the Y axis 0 line. Non-homoscedasticity is shown when points form a funnel or other shape showing variance differs as one moves along theY axis.

- **Additivity**. Likewise, regression does not account for interaction effects, although interaction terms (usually products of standardized independents) may be created as additional variables in the analysis. As in the case of adding nonlinear transforms, adding interaction terms runs the danger of overfitting the model to what are, in fact, chance variations in the data. Such terms should be added only when there are theoretical reasons for doing so. That is, significant but small interaction effects from interaction terms not added on a theoretical basis may be artifacts of overfitting. Such artifacts are unlikely to be replicable on other datasets.

- **Homoscedasticity**: The researcher should test to assure that the residuals are dispersed randomly throughout the range of the estimated dependent. Put another way, the variance of residual error should be constant for all values of the independent(s). If not, separate models may be required for the different ranges. Also, when the homoscedasticity assumption is violated "conventionally computed confidence intervals and conventional t-tests for OLS estimators can no longer be justified" (Berry, 1993: 81).

  Nonconstant error variance can be observed by requesting a simple residual plot (see above). A homoscedastic model will display a cloud of dots, whereas lack of homoscedasticity will be characterized by a pattern such as a funnel shape, indicating greater error as the dependent increases. Nonconstant error variance can indicate the need to respecify the model to include omitted independent variables.

  Lack of homoscedasticity may mean (1) there is an interaction effect between a measured independent variable and an unmeasured independent variable not in the model; or (2) that some independent variables are skewed while others are not. One method of dealing with hetereoscedasticity is to selecte the *weighted least squares* regression option. This causes cases with smaller residuals to be weighted more in calculating the b coefficients.

  - **No outliers**. Outliers are a form of violation of homoscedasticity. Detected in the analysis of residuals and leverage statistics, these are cases representing high residuals (errors) which are clear exceptions to the regression explanation. The set of outliers may suggest/require a separate explanation. Some computer programs allow an option of listing outliers directly, or there may be a "casewise plot" option which shows cases more than 2 s.d. from the estimate. To deal with outliers, the researcher may remove them from analysis and seek to explain them on a separate basis, or transforms may be used which tend to "pull in" outliers. These include the square root, logarithmic, and inverse (x = 1/x) transforms.

- **Reliability**: To the extent there is random error in measurement of the variables, the regression coefficients will be attenuated. To the extent there is systematic error in the measurement of the variables, the regression coefficients will be simply wrong. (In contrast to OLS regression, structural equation modeling involves explicit modeling of measurement error, resulting in coefficients which, unlike regression coefficients, are unbiased by measurement error.) Note measurement error terms are not to be confused with residual error of estimate, discussed below.

- **Normally distributed residual error**: Error, represented by the residuals, should be normally distributed for each set of values of the independents. A histogram of standardized residuals should show a roughly normal curve. An alternative for the same purpose is the normal probability plot, with the observed cumulative probabilities of

occurrence of the standardized residuals on the Y axis and of expected normal probabilities of occurrence on the X axis, such that a 45-degree line will appear when observed conforms to normally expected. The *central limit theorem* assumres that even when error is not normally distributed, when sample size is large, the sampling distribution of the b coefficient will still be normal. Therefore violations of this assumption usually have little or no impact on substantive conclusions for large samples, but when sample size is small, tests of normality are important.

- **Population error is uncorrelated with each of the independents**). This is the "assumption of mean independence": that the mean error is independent of the x independent variables. This is a critical regression assumption which, when violated, may lead to substantive misinterpretation of output.

  The (population) error term, which is the difference between the actual values of the dependent and those estimated by the population regression equation, should be uncorrelated with each of the independent variables. Since the population regression line is not known for sample data, the assumption must be assessed by theory. Specifically, one must be confident that the dependent is not also a cause of one or more of the independents, and that the variables not included in the equation are not causes of Y and correlated with the variables which are included. Either circumstance would violate the assumption of uncorrelated error. One common type of correlated error occurs due to *selection bias* with regard to membership in the independent variable "group" (representing membership in a treatment vs. a comparison group): measured factors such as gender, race, education, etc., may cause differential selection into the two groups and also can be correlated with the dependent variable. When there is correlated error, conventional computation of standard deviations, t-tests, and significance are biased and cannot be used validly.

  Note that *residual error* -- the difference between observed values and those estimated by the sample regression equation -- will always be uncorrelated and therefore the lack of correlation of the residuals with the independents is not a valid test of this assumption.

  > **Two-Stage Least Squares (2SLS),** discussed separately, is designed to extend the regression model to situations where non-recursivity is introduced because the researcher must assume the correlations of some error terms are not 0. It can be used, for instance, to test for selection bias. Click here for the separate discussion.

- **Independent observations (absence of autocorrelation) leading to uncorrelated error terms.** Current values should not be correlated with previous values in a data series. This is often a problem with time series data, where many variables tend to increment over time such that knowing the value of the current observation helps one estimate the value of the previous observation. Spatial autocorrelation can also be a problem when units of analysis are geographic units and knowing the value for a given area helps one estimate the value of the adjacent area. That is, each observation should be independent of each other observation if the error terms are not to be correlated, which would in turn lead to biased estimates of standard deviations and significance.

  - The **Durbin-Watson coefficient, d,** tests for autocorrelation. The value of d ranges from 0 to 4. Values close to 0 indicate extreme positive autocorrelation; close to 4 indicates extreme negative autocorrelation; and close to 2 indicates no serial autocorrelation. As a rule of thumb, d should be between 1.5 and 2.5 to indicate independence of observations. Positive autocorrelation means standard errors of the b coefficients are too small. Negative autocorrelation means standard errors are too large.

    Alternatively, the d value has an association p probability value for various significance cutoffs (ex., .05). For a given level of significance such as .05, there is an upper and a lower d value limit. If the computed Durbin-Watson d value for a given series is more than the upper limit, the null hypothesis of no autocorrelation is not rejected and it is assumed that errors are serially uncorrelated. If the computed d value is less than the lower limit, the null hypothesis is rejected and it is assumed that errors are serially correlated. If the computed value is in-between the two limits, the result is inconclusive. In SPSS, one can obtain the Durbin-Watson coefficient for a set of residuals by opening the syntax window and running the command,

FIT RES_1, assuming the residual variable is named RES_1.

For a graphical test of serial independence, a plot of residuals on the Y axis against the sequence of cases (the caseid variable) on the X axis should show no pattern, indicating independence of errors.

When autocorrelation is present, one may choose to use **generalized least-squares (GLS)** estimation rather than the usual ordinary least-squares (OLS). In iteration 0 of GLS, the estimated OLS residuals are used to estimate the error covariance matrix. Then in iteration 1, GLS estimation minimizes the sum of squares of the residuals weighted by the inverse of the sample covariance matrix.

- Validity. As with all procedures, regression assumes measures are valid.

- **Mean population error is zero**: The mean of the (population) error term (see above) should be zero. Since the population regression line is not known for sample data, the assumption must be assessed by analysis of nonresponse (see the section on survey research). Specifically, one must be confident that there is no selection bias causing certain subpopulations to be over- or under-represented. Note that *mean residual error* is always zero and thus is not a valid test of this assumption.

- **Random sampling** is not required for regression, but if used with enumeration data for an entire population, then significance tests are not relevant. When used with non-random sample data, significance tests would be relevant but unfortunately cannot be reliable and thus are not appropriate. Nonetheless, social scientists commonly use significance tests with non-random data due to their utility as an arbitrary decision criterion.

Example of SPSS Regression Output

- SPSS Regression Output

Frequently Asked Questions

- What is the logic behind the calculation of regression coefficients in multiple regression?
- All I want is a simple scatterplot with a regression line. Why won't SPSS give it to me?
- How big a sample size do I need to do multiple regression?
- Can R-squared be interpreted as the percent of the cases xplained?
- When may ordinal data be used in regression?
- How do I use dummy variables in regression?
- What is "attenuation" in the context of regression?
- Is multicollinearity only relevant if there are significant findings?
- What can be done to handle multicollinearity?
- How does "corresponding regressions" aid causal analysis?
- How does stepwise multiple regression relate to multicollinearity?
- What are forward inclusion and backward elimination in stepwise regression?
- What is part correlation in regression output?
- Can regression be used in place of ANOVA for analysis of categorical independents affecting an interval dependent?
- Does regression analysis require uncorrelated independent variables?
- How can you test the significance of the difference between two R-squareds?
- How do I compare b coefficients after I compute a model with the same variables for two subgroups of my sample?
- How do I compare regression results obtained for one group of subjects to results obtained in another group, assuming the same variables were used in each regression model?

- **What is the logic behind the calculation of regression coefficients in multiple regression?**

  Let Y be the dependent and X_1 and X_2 be the two independents. Multiple regression first regresses X_1 on X_2 (and other independents, if there were any) and sets aside these residuals, which represent the unique variance in X_1 uncorrelated with other independent variables. The process is repeated for the regression of X_2 on X_1, also setting aside these residuals. As a last step, Y is regressed on the sets of residuals for each of the independents. The resulting b coefficients are the partial regression coefficients which reflect the unique association of each independent with the Y variable. The interpretation of the intercept, c, is the same in bivariate and multiple regression, whereas note that the b coefficients differ: they are simple coefficients for bivariate regression but are partial coefficients for multiple regression.

- **All I want is a simple scatterplot with a regression line. Why won't SPSS give it to me?**

  SPSS has hidden it. When you choose Statistics/Regression/Linear and click on Plots, you will find a choice for a dependent axis but not an independent axis and you can't get the plot you want. Instead, ignore the Plots button and do the regression anyway. Then in the output, click on Graphs on the menu bar at the top and choose Scatter to get the Scatterplot dialog box. Choose the type you want (probably "Simple") then click Define to get another Scatterplot dialog box for the type you picked. This box will have places to enter your independent and dependent variables, then click OK to create the plot. Once the plot appears, double-click on it to bring up the Chart Editor, where you select Chart, then Options, then check Fit Line in the dialog box. The result will be the plot you want.

- **How big a sample size do I need to do multiple regression?**

  According to Tabachnick and Fidell (2001: 117), a rule of thumb for testing b coefficients is to have N >= 104 + m, where m = number of independent variables. If you are using stepwise regression, N >= 40m is a rule of thumb since stepwise methods can train to noise too easily and not generalize in a smaller dataset. A rule of thumb for testing R-square is N >= 50 + 8m. Where m>=N, regression gives a meaningless solution with R=square = 1.0. In general, you need a larger N when the dependent variable is skewed; you are seeking to test small effect sizes (rule of thumb: N>= (8/f2) + (m-1), where f2 = .01, .15, and .35 for small medium and large effect sizes); where there is more measurement error in the independent variables; and of course if your design calls for cross-validating training data to test data.

- **Can R-squared be interpreted as the percent of the cases explained?**

  No. Sometimes one reads an author saying that because R^2 is .40, therefore 40% of the cases will be correctly predicted by the regression equation and 60% not. However, the coefficient of determination has no such goodness-of-classification-interpretation. Rather R^2 decomposes the total variation.

- **When may ordinal data be used in regression?**

  Technically, never.

As an independent: The regression model makes no distributional assumptions about the independents, which may be discrete variables as long as other regression assumptions are met. The discreteness of ordinal variables is thus not a problem, but do ordinal variables approach intervalness? Ordinal variables must be interpreted with great care when there are known large violations of intervalness, such as where it is known that rankings obscure large gaps between, say the top three ranks and all the others. In most cases, however, methodologists simply use a rule-of-thumb that there must be a certain minimum number of classes in the ordinal independent (Achen, 1991, argues for at least 5; Berry (1993: 47) states five or fewer is "clearly inappropriate"; others have insisted on 7 or more). However, it must be noted that use of 5-point Likert scales in regression is extremely common in the literature.

As a dependent: Ordinal dependents are more problematic because their discreteness violates the regression assumptions of normal distribution of error with constant variance. A conservative method is to test to see if there are significant differences in the regression equation when computed separately for each value class of the ordinal dependent. If the independents seem to operate equally across each of the ordinal levels of the dependent, then use of an ordinal dependent is considered acceptable. The more liberal and much more common approach is to allow use of ordinal dependents as long as the number of response categories is not very small (at least 5 or 7, see above) and the responses are not highly concentrated in a very small number of response categories.

- **How do I use dummy variables in regression?**

Regression assumes interval data, but dichotomies may be considered a special case of intervalness. Nominal and ordinal categories can be transformed into sets of dichotomies, called dummy variables. To prevent perfect multicollinearity, one category must be left out. For instance, for the nominal variable "Region" we may create a set of dummy variables called East, West, and South, leaving out North.

Three considerations govern which category to leave out. Since the b coefficients for dummy variables will reflect changes in the dependent with respect to the reference group (which is the left-out group), it is best if the reference group is clearly defined. Thus leaving out the "Other" or "Miscellaneous" category is not a good idea since the reference comparisons will be unclear, though leaving out "North" in the example above would be acceptable since the reference is well defined. Second, the left-out reference group should not be one with only a small number of cases, as that will not lead to stable reference comparisons. Third, some researchers prefer to leave out a "middle" category when transforming ordinal categories into dummy variables, feeling that reference comparisons with median groups are better than comparisons with extremes.

Regression coefficients should be assessed for the entire set of dummy variables for an original variable like "Region" (as opposed to separate t-tests for b coefficients as is done for interval variables). For a regression model in which all the independents are dummies for one original ordinal or nominal variable, the test is the F test for R-squared. Otherwise the appropriate test is the F test for the difference of R-squareds for the model with the set of dummies and the model without the set.

$$F = [(R_2^2 - R_1^2)/(k_2 - k_1)]/[(1-R_2^2)/(n - k_2 -1)]$$

where the subscripts refer to model 1 (without the set of dummies) and model 2 (with the set of dummies); where k refers to the number of independent variables in the model; n is the sample size; and degrees of freedom for the F test is $k_2 - k_1$ and $N - k_2 - 1$).

There are three methods of coding dummy variables. Coding greatly affects the magnitude and meaning of the b and beta coefficients, but not their significance. Coding does not affect the R-squared for the model or the significance of R-squared, as long as all dummy variables save the reference category are included in the model.

1. Binary coding, also called *indicator coding* or *dummy coding*, is by far the most common and makes comparisons in relation to the omitted reference group. In *binary coding*, if a unit is in the East it

would be coded 1 on a variable called "East" and 0 on "West" and "South", for instance. If the resulting b coefficient is, say 2.1, this means that being in the East causes the response (dependent) variable to increase by 2.1 units compared to the unit being in the North, which is the reference (left-out) category. This implies that if North is included in the model and East is left out, then the b coefficient for North also will be 2.1. A positive b coefficient for any included group means it scored higher on the response variable than did the reference group, or if negative, then lower. A significant b coefficient for any included group means that group is significantly different on the response variable from the reference group.

In general, the b coefficients are the distances from the dummy values to the reference value, controlling for other variables in the equation, and the distance from the reference category to the other dummy variables will be the same in a model in which the reference (omitted) categories are switched. Another implication is that the distance from one included dummy value to another included value (ex., from East to West in the example in which North is the omitted reference category) is simply the difference in their b coefficients. Thus if the b coefficient for West is 1.6, then we may say that the effect of East is .5 units more (2.1 - 1.6 = .5) than the West effect, where the effect is still gauged in terms of unit increases in the dependent variable compared to being in the North. For "Region," assuming "North" is the reference category and education level is the dependent, a b of -1.5 for the dummy "South" means that the expected education level for the South is 1.5 years less than the average of "North" respondents.

Some textbooks say the b coefficient for a dummy variable is the difference in means between the two values of the dummy (0,1) variable. This is true only if the variable is a dichotomy. In general, the b coefficient for a give dummy variable is the difference in means between the given dummy variable and omitted reference dummy variable. For dichotomies, there will be only one given dummy variable ant the other value will be the omitted reference category and so it is a special case in which the b coefficient is the difference in means between the two values of the dummy variable.

In an experimental context, the omitted reference group would ordinarily be the control group.

2. Effect coding, also called *deviation coding*, makes comparisons in relation to the grand mean of all subgroups. The researcher picks a category, such "South" from the set "Region", and codes it 1. Another category, such as "West", is coded -1, to indicate it is the reference category. All remaining groups are coded 0, and they will not affect outcomes if group sizes are equal, and even when sizes are unequal, effects will be small. One group is left out, as usual.

   Given this effect coding and education level as the dependent, a b of -1.5 for the dummy "South" means that the expected education level for the South is 1.5 years less than the unweighted mean of the expected values for all subgroups. That is, binary coding interprets b for the dummy category (South) relative to the reference group (the left-out category), effects coding interprets it relative to the entire set of groups. A positive b coefficient for any included group (other than the -1 group) means it scored higher on the response variable than the grand mean for all subgroups, or if negative, then lower. A significant b coefficient for any included group (other than the -1 group) means that group is significantly different on the response variable from the grand mean. Under effect coding there is no comparison between the group coded -1 and the grand mean. nc

3. Contrast coding, also called *orthogonal coding*, lets the researcher specify just which constrasts are to be tested. Contrast coding allows the researcher to establish clusters of categories and contrast them. For instance, in the dummy variable set which includes occupational types (managers, white-collar, skilled, service, and laborer), manager/white-collar might be considered one cluster and skilled/service/laborer might be considered another.

   To compare the first cluster with the second, the cluster of interest (managers and white-collar) would

thus be coded +.5 each (1 divided by the 2 categories in the cluster), and the other categories of the reference cluster as -.33 each (-1 divided by the 3 categories). Contrast code(s) will sum to zero across all categories. To contrast managers v. white-collar only, considering managers as the category of interest (coded +1), white-collar the reference category (coded -1), and all others as the third cluster (coded 0). The *group contrast* is the b coefficient times ([nint + nref]/[(nint)*(nref)], where n is the number of categories for the cluster of categories of interest (int) or the reference cluster (ref).

A significant b coefficient means the variables or clusters of variables being contrasted are significantly different on the response variables. Under contrast coding, the b coefficients do not have a clear interpretation in terms of group means on the response variable.

- **What is "attenuation" in the context of regression?**

  Regression (b) coefficients may be corrected for attenuation. The greater the variance of measurement error, the more the regression coefficient is biased toward zero. Severe error noise in the data will lead to severe underestimation of the true regression coefficients, for example.The amount of underestimation is *attenuation*. In the case of bivariate regression of Y on X, the attenuation is estimated as the ratio of the variance of X to the sum of the variance in X plus the error variance. This ratio is the *attenuation coefficient*. The lower the reliability of a variable, the more the attenuation. Because attenuation leads to underestimates of regression coefficients it also undermines the meaningfulness of significance tests of regression models, reducing statistical power and increasing the likelihood of Type II errors.

  See Fuller (1987), who for a not untypical data set estimated attenuation coefficients of .98 for gender, .88 for education level, and .58 for poverty status. That is, attenuation is a non-trivial problem which can lead to serious underestimation of regression coefficients. The variance of the residuals is the estimate of error variance, assuming all relevant variables are in the equation and all irrelevant variables are omitted.

- **Is multicollinearity only relevant if there are significant findings?**

  It is sometimes claimed that multicollinearity does not matter if research comes to nul findings, with no significant results. The argument is that multicollinearity undermines the ability to rank the importance of independent variables, but if all independent variables are non-significant, then such ranking is moot and multicollinearity does not matter. This is false reasoning. Multicollinearity increases the standard errors of the b coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0, whereas without multicollinearity and with lower standard errors, these same coefficients might have been found to be significant and the researcher may not have come to nul findings in the first place.

- **What can be done to handle multicollinearity?**
  1. Increasing the sample size is a common first step since when sample size is increased, standard error decreases (all other things equal). This partially offsets the problem that high multicollinearity leads to high standard errors of the b and beta coefficients.
  2. Use **centering**: transform the offending independents by subtracting the mean from each case. The resulting centered data may well display considerably lower multicollinearity. You should have a theoretical justification for this consistent with the fact that a zero b coefficient will now correspond to the independent being at its mean, not at zero, and interpretations of b and beta must be changed accordingly.
  3. Combine variables into a composite variable. This requires there be some theory which justifies this conceptually.
  4. Remove the most intercorrelated variable(s) from analysis. This method is misguided if the variables were there due to the theory of the model, which they should have been.
  5. Drop the intercorrelated variables from analysis but substitute their crossproduct as an interaction term, or in some other way combine the intercorrelated variables. This is equivalent to respecifying the model by conceptualizing the correlated variables as indicators of a single latent variable. Note: if a correlated variable is a dummy variable, other dummies in that set should also be included in the combined variable in order to keep the set of dummies conceptually together.
  6. Leave one intercorrelated variable as is but then remove the variance in its covariates by regressing them on

that variable and using the residuals.

7. Assign the common variance to each of the covariates by some probably arbitrary procedure.
8. Treat the common variance as a separate variable and decontaminate each covariate by regressing them on the others and using the residuals. That is, analyze the common variance as a separate variable.
9. Use orthogonal principal components factor analysis, then use the factors as the independents.
10. **Ridge regression** is an attempt to deal with multicollinearity through use of a form of biased estimation in place of OLS. The method requires setting an arbitrary "ridge constant" which is used to produce estimated regression coefficients with lower computed standard errors. However, because picking the ridge constant requires knowledge of the unknown population coefficients one is trying to estimate, Fox (1991: 20) and others recommend against its use in most cases. SPSS has no ridge regression procedure, but its macro library has the macro ridge_regression.sps.

- **How does "corresponding regressions" aid causal analysis?**

Chambers (1986) observed that causality could be inferred from the correspondence of variances in dependent variables. Using simulation, he demonstrated that high values of a dependent arise from high values of the independents, low values from low values, but moderate values of the dependent may arise from many levels of the independents as high and low independent values cancel each other out. He also showed that restricting the independents to moderate values led generally to moderate values of the dependent. Based on these observations, Chambers showed that the variance of dependent variables corresponding to cases with mid-range independents is lower than the variance of independent variables corresponding to cases with mid-range dependent variables. This assymetry is used to infer causal direction.

The method of causal inference through corresponding regressions was subsequently set out by Chambers (1991). Consider bivariate regression of y on x, but where there is uncertainty about whether the causal direction should not be the opposite. In corresponding regressions, y is regressed on x, and the absolute deviations (predicted minus actual values of y) are noted as a measure of the extremity of prediction errors. Next the deviations of the x values from the mean of x are taken to give a measure of the extremity of the predictor values. The two columns of deviations are correlated, giving the deviation correlation for y, labeled rde(y). The deviation correlation will be negative, since when predictor values are extreme, errors should be less since high values of the predictor lead to high values of the dependent, and low values to low values. The regression is then repeated for the regression of x on y, giving rde(x).

When the real independent variable serves as a predictor, there should be a higher correlation than when the real dependent serves as predictor. That is, the rde() value is higher when the real independent serves as the predictor. This is because mid-range predictor values (as measured by low extremity of predictor variables) should be associated with mid-range dependent values (as measured by the extremity of errors) only when the true independent is used as the predictor of the true dependent. **Chambers' D** is rde(y) - rde(x). When the true independent is x and the true dependent is y, D will be negative. That is, only if x is the true independent and y is the true dependent will rde(y) be more negative than rde(x), and D will have a negative value after subtraction. If it does not, Chambers recommends assuming no correlation of the two variables (1991: 12).

Assumptions of corresponding regressions

1. *Bivariate causality.*Corresponding regressions investigates the causal relationship between two variables. Of course, additional unmeasured variables may be causes of the two measured variables, with the effect of these unmeasured variables found in the regression error term. Chambers' simulations showed D to be efficient for hierarchical models in which the true independent was caused by a chain of prior variables.
2. *Sample size.* Chambers' simulations show D to require at least a moderate sample size (ex., 50 or more).
3. *Correlation.* Chambers' simulations show D to be nearly 100% accurate when n>50 and the correlation of the independent with the dependent was in the range .2 to .9. The method is not appropriate when the correlation of the independent and dependent is outside this range.
4. *Additivity.* Chambers' simulations show the efficiency of D to be reduced for multiplicative models.

Rather, corresponding regressions assumes the independent and error term additively determine the dependent.

Note that corresponding regressions is a controversial method not yet widely accepted and applied in the social science literature.

- **How does stepwise multiple regression relate to multicollinearity?**
  Stepwise procedures select the most correlated independent first, remove the variance in the dependent, then select the second independent which most correlates with the remaining variance in the dependent, and so on until selection of an additional independent does not increase the R-squared by a significant amount (usually signif = .05). While stepwise regression uses a meaningful criterion for selecting variables, it does not assure that the selected variables will not display high multicollinearity (high intercorrelation).

- **What are forward inclusion and backward elimination in stepwise regression?**
  Forward inclusion is the option, usually the default in computer programs, of entering the best variable first, then the second partially for the first, and so on. Backward elimination is the alternative option of starting with all variables in the equation, then eliminating independents one at a time until such an elimination makes a significant difference in R-squared. Forward inclusion provides a rationale for intelligent but automated ordering of variables but it will miss independents which exhibit *suppressor* effects -- variables whose significant relationship to the dependent is only apparent when other variables are controlled. For instance, a variable which affects the dependent positively through one intervening variable and negatively through another may seem to have no significant relationship with the dependent and may not be included in the model under forward inclusion. If suppression is suspected, backward elimination should be chosen as the stepwise option.

- **What is part correlation in regression output?**
  This is discussed in the section on partial correlation.

- **Can regression be used in place of ANOVA for analysis of categorical independents affecting an interval dependent?**
  Yes. One would have to use dummy variables for the independents and would have to include explicit crossproduct interaction terms. When using dummy variables one would have to leave out one of the values of each categorical independent to prevent overdetermination. When all interaction terms are included, the F value for the regression equation will be equal to the F value of the Explained row in the ANOVA output.

  Note that ANOVA is not interchangeable with regression for two reasons: (1) ANOVA cannot handle continuous variables as it is a grouped procedure. While continuous variables can be coded into categories, this loses information and attenuates correlation; and (2) ANOVA normally requires approximately equal n's in each group formed by the intersection of the independent variables. Equal group sizes is equivalent to orthogonality among the independent variables. Regression allows correlation among the IVs (up to a point, lower than multicollinearity) and thus is more suitable to non-experimental data. Methods exist in ANOVA to adjust for unequal n's but all are problematic.

- **Does regression analysis require uncorrelated independent variables?**
  This is occasionally said, but misleadingly. Regression analysis assumes uncorrelated error terms, but not uncorrelated independents. However, it is true that the less correlation of the independents, the less multicollinearity will be a problem. so in this sense the statement is true.

- **How can you test the significance of the difference between two R-squareds?**
  For instance, for the F test of differences between two regression models where one includes interaction effects and one does not, use an F test:

$$F = [(R_2^2 - R_1^2)/(k_2 - k_1)]/[(1-R_2^2)/(n - k_2 -1)]$$

Where

R_2^2 = R-square for the second model (ex., one with interactions or with an added independent)

R_1^2 = R-square for the first, restricted model (ex., without interactions or without an added independent)

n = total sample size

k_2 = number of predictors in the second model

k_1 = number of predictors in the first, restricted model

F has $(k_2 - k_1)$ and $(n - k_2 - 1)$ degrees of freedom and tests the null hypothesis that the R^2 increment between the two models is not significantly different from zero.

- **How do I compare b coefficients after I compute a model with the same variables for two subgroups of my sample?**

  A t-test of difference of b coefficients for separate subgroup regressions is available. See Hardy (1993: 52) for discussion.

- **How do I compare regression results obtained for one group of subjects to results obtained in another group, assuming the same variables were used in each regression model?**

  The best procedure is the **Chow test**. Combine data for the two groups into one dataset, adding a variable which indicates from which group each row (case) came. Then create interaction terms by multiplying this grouping variable times each other independent in turn, creating as many interaction terms as there are other independents. The researcher uses the ENTER method and specifies the original dependents as the first block, then specifies all the interaction terms as the second block. In the output, under "Model Summary" in SPSS, the "Sig. F Change" column tests the null hypothesis that the regressions for the two groups are equal. An alternative suggested by SPSS Senior Statistician David Nichols is to do a similar analysis in MANOVA, which will not require creating the grouping dummy variable and all the interaction terms:

  ```
  MANOVA Y BY GROUP(1,K) WITH X1 X2 XJ
  /ANALYSIS Y
  /DESIGN=X1 X2 XJ GROUP+GROUP BY X1+GROUP BY X2+GROUP BY XJ.
  ```

- **What do I do if I have more observations on my independents than on my dependents?**

  Biased estimation of b coefficients is likely to occur when using regression with censored, sample selected, or truncated data. Censored data is where the researcher has all data for the independents but has data for the dependent only if some criterion related to the dependent is met (ex., if the dependent is above a threshold value). Sample selected data is the same thing but where the criterion relates to a third variable (ex., there is data on the dependent if the third variable is above a threshold). Truncated data is where there is data for the independents only if there is data for the dependent. Breen (1996) recommends the use of maximum likelihood Tobit estimates of linear model coefficients in preference to standard OLS estimates, because Tobit, makes expected values of the dependent conditional on the probability of censoring or sample selection. The LIMDEP and SHAZAM statistics packages support Tobit, as does SAS using SAS/IML software. Often the censored or sample-selected data have to do with time as a criterion (the dependent is observed only after a certain time period). This type of censored data is handled in SPSS using Cox regression (also a maximum likelihood method), Kaplan-Meier Survival analysis, or the Life Tables procedure.

- **What do I do if I am measuring the same independent at both the individual and group level? What is contextual analysis in regression?**

  Iversen (1991) addresses how regression may be used to investigate how a variable operates at the individual and group levels as, for instance, to separate the effects on performance of individual ability, team ability, and the interaction of the two. He critiques the usual regression "absolute effects" model in which, for this example, the individual-level measure would be an ability score and the group-level measure would be the mean ability score of all members on the individual's team. Iverson shows how multicollinearity in such

models, particularly when group/individual interaction terms are added, can make it difficult or impossible to separate group effects from individual-level measures and leads to unreliable analysis.

Instead Iverson proposes a "relative effects model" in which the individual-level measure would be, for this example, the individual's ability score minus the group (team) mean, and the group-level measure would be the group team mean minus the overall mean on ability for all teams. This transformation, which must be warranted by the theory of the model being assessed, usually eliminates or greatly reduces the multicollinearity problem. In the relative effects model, one would then regress performance on the relative individual ability measures, employing a separate regression for each team. The constant is the value of performance when the individual ability is the same as the team mean (not zero, as in the absolute effects model). If the b coefficients vary from team to team, this indicates a group effect.

To investigate the group effect using the single-equation method one regresses performance on the relative individual, group, and interaction variables, generating coefficients corresponding to the individual, group, and interaction effects. (Iverson also describes a separate equation method which generates the same estimates but the single-equation method usually has smaller standard errors). The standardized coefficients (beta weights) in this regression allow comparison of the relative importance of the individual, group, and interaction effects. This comparison does not suffer from multicollinearity as the relative effects transformations leave the variables with little or no correlation in most cases. Iverson (ppp. 64-66) also describes an alternative of partitioning the sums of squares to assess individual vs. group vs. interaction effects.

The models Iverson discusses can be done in SPSS or SAS, but one must compute the relative individual, group, and interaction variables manually. This can become tedious or nearly impossible in large models. Consequently, various packages for contextual analysis have been created, including GENMOD (Population Studies Center, University of Michigan), ML3 (Multilevel Models Project, Institute of Education, University of London), and the one most popular, HLM (see Bryk et al., 1988). Iverson briefly mentions these packages but provides no discussion of the steps involved in their use.

- **How do I test to see what effect a quadratic or other nonlinear term makes in my regression model?**
  Add the quadratic term (ex., income-squared) as an additional independent in your model. Researchers often center their data (subtract the mean) prior to applying the quadratic transformation, thereby giving them what is called an *orthogonal polynomial*. Compare the resulting R^2 with the R^2 for the linear model without the quadratic term, using the usual test for R^2 difference.

- **When testing for interactions, is there a strategy alternative to adding multiplicative interaction terms to the equation and testing for R^2 increments?**
  There are two other strategies. In one, the two variables suspected of interacting are dichotomized around their median values (or some other criterion), then a traditional 2x2 ANOVA is conducted, giving main and interaction effects. In the second strategy, one of the interacting independents is dichotomized, dividing the sample into two groups. Separate regressions of the dependent on the other interacting independent are conducted, and the b coefficients are compared. To the extent they are similar, there is no interaction effect. However, because dichotomization loses information, these methods are generally less preferred. See Jaccard, Turrisi, and Wan (1991: 48-49) for further discussion of these alternatives.

- **What is "smoothing" in regression and how does it relate to nonlinear regression?**
  Smoothing is fitting a nonlinear line through the points on a scatterplot. Nonlinear regression is adding polynomial terms to the regression equation, or it is nonparametric regression, discussed below. Fox (2000a) covers various types of smoothing and nonparametric regression. Fox distinguishes these types of smoothing:

  - *Binning*. The independent (x) variable is divided into non-overlapping ranges called bins and the regression line is calculated separately for each bin, resulting in a a series of regression lines

connected in stair-type steps. Usually the bin cutting points are set so as to have equal numbers of cases in each bin. The number of bins must be small enough so that the regression for any given bin is based on a sufficient number of cases as to still be significant.

- *Local Averages.* This method is similar to binning, but a bin (bandwith, window) is moved across the independent (x_ variable range and regressions are calculated at a large number of equally spaced data points. This method typically involves "boundary bias" -- flattening of the regression line at the left and right edges of the distribution. Note that when the x-variable is time, local averages are called *moving averages*, a common technique in time series analysis.

- *Kernel Estimation.* Kernel estimation is local averaging with weighting, such that cases nearer to the center of the bin are weighted more. Kernel weighting is usually of the normal distribution type, but other weighting distributions are possible. Because outliers can affect kernel estimates radically, weighting may be done in a two-step process under which cases with high residuals (outliers) are downweighted in the second stage. Biweighting (bisquare) and Huber weighting are types of such downweighting (see Fox, 2000a: 40-41). Kernel estimation will result in a smoother line than local averaging but will still display boundary bias.

- *Local Polynomial Regression.* Local polynomial regression is kernel estimation, except that the regression line fitted in each bin are polynomial ones. Local polynomial regression has lower bias (smaller mean squared error) than kernel estimation because it can account for nonlinear data relationships. Average squared error (ASE) is a measure of the accuracy of local polynomial regression. The most common form of local polynomial regression is *Loess* (LOcal regrESSion). See below.

  The researcher can set the level of exponentiation (including 1 = the linear case), but cubic polynomial fitting is typical. Thus, for simple one-independent variable models, let x_0 be the value of x at the bin focal point, and let x_i be the value of x at any of i other points within the bin. In the cubic case, the polynomial regression equation would be $y_i = b_1(x_i - x_0) + b_2(x_i - x_0)^2 + b_3(x_i - x_0)^3 + c$.

  The span, s, for the bandwidth can also be set by the researcher. One method is simply visual trial and error using various values of s, seeking the smallest s which still generates a smooth curve.

- *Regression Splines.* Cubic regression splines operate similar to local polynomial regression, but a constraint is imposed that the regression line in a given bin must join to the start of the regression line in the next bin, thereby avoiding discontinuities in the curve, albeit by increasing error a bit.

- *Smoothing Splines.* Smoothing splines are a more complex refinement to regression splines but they do not generalize to multiple independents and thus are less used. See Fox (2000a: 67-69).

- **What is nonparametric regression for nonlinear relationships?**
  "Nonparametric regression" (smoothing) refers to a family of techniques which relax OLS or GLM regression's assumption that the term on the left-hand side of the regression equation (the dependent variable in OLS, the logit of the dependent in GLM) is a linear function of the terms on the right-hand side. Nonparametric regression fits a smoothed curve to the data scatterplot rather than a straight line. Comparing a conventional regression model with the corresponding nonparametric one would be a test of nonlinearity.

  For handling nonlinear relationships in a regression context, nonparametric regression is now considered preferable to simply adding polynomial terms to the regression equation (as is done, for instance, in SPSS through Analyze, Regression, Nonlinear menu choice). Nonparametric regression methods allow the data to influence the shape (curve) of a regression line. Note that this means that nonparametric regression is usually an atheoretical method, not involving positing the model in advance, but instead deriving it from the

data. Consequently, fitting a curve to noise in the data is a critical concern in nonparametric regression. Nonparametric regression is treated in Fan and Gijbels (1996) and Fox (2000b), who covers local polynomial multiple regression, additive regression models, projection-pursuit regression, regression trees, and GLM nonparametric regression.

Local polynomial multiple regression makes the dependent variable a single nonlinear function of the independent variables. Local regression fits a regression surface not for all the data points as in traditional regression, but for the data points in a "neighborhood." Researchers determine the "smoothing parameter," which is a specified percentage of the sample size, and neighborhoods are the points within the corresponding radius. In the loess method, weighted least squares is used to fit the regression surface for each neighborhood, with data points in a neighborhood weighted in a smooth decreasing inverse function of their distance from the center of the neighborhood. Alternative to this nearest-neighbor smoothing, one may define bands instead of neighborhood spans, with bandwidths being segments of the range of the independent variable(s). The fitting of surfaces to neighborhoods may be done at a sample of points in predictor space, or at all points. Regardless, the surfaces are then blended together to form the curved line or curved surface characteristic of nonparametric regression. SAS implements local regression starting in Version 8 in its proc loess procedure. As of version 10, SPSS does not directly implement nonparametric regression though its website does provide an java applet demo. See Fox (2000b: 8-26).

Problems of local regression. Fox (2000b: 20) refers to "the curse of dimensionality" in local regression, noting that as the number of predictor variables increases, the number of data points in the local neighborhood of a focal point tends to decline rapidly. This means that to obtain a given percentage of data points, the smoothing parameter radius must become less and less local. Other problems of local regression include (1) its post-hoc atheoretical approach to defining the regression curve; (2) the fact that dynamic inference from the b coefficients is no longer possible due to nonlinearity, requiring graphical inference instead; and (3) graphical display becomes difficult to comprehend when more than three independent variables are in the model (Fox, 2000b: 26, recommends coplots as the best display alternative).

Additive regression models allow the dependent variable to be the additive sum of nonlinear functions which are different for each of the independent variables. This means that the dependent variable equals the sum of a series of two-dimensional partial regressions. For the dependent y and each independent x, one can predict adjusted y as a local regression function of x. The adjustment has to control y for other independents in the equation. An iterative method called backfitting simultaneously solves the nonlinear functions for each independent (x) term, and the dependent is the additive sum of these terms. See Fox (2000b: 27-37).

Note it is also possible to have a *semi-parametric regression model* in which some of the independent variables have nonparametric functions as described above, while others have conventional regression coefficients. In particular, a semi-parametric model would be appropriate if dummy variable terms were present: dummy variables would be entered as linear terms. Additive models have the same problems of interpretation as local regression.

Projection-pursuit regression first reduces attribute space by creating latent variables which are regression functions of the raw independent variables, then makes the dependent the additive sum of nonlinear functions which are different for each of these latent variables. A purpose of projection-pursuit regression is that by reducing the number of variables in local regression and by making the dependent an additive function of a series of bivariate partial regressions, the "curse of dimensionality" problem mentioned above is mitigated. The price paid, however, is that, as Fox notes, "arbitrary linear combinations of predictors do not usually correspond to substantively meaningful variables" and difficulty in interpreting the resulting nonparametric regression is multiplied. At least with additive regression models, for instance, one can interpret the partial regression coefficient signs as indications of the direction of effect of individual predictor variables. See Fox (2000b: 37-47).

Regression trees employ successive binary divisions of predictor attribute space, making the dependent variable a function of a binning and averaging process. Also called the AID (automatic interaction detection)

method, regression trees are classification trees for continuous data. There are several different algorithms for creating regression trees, but they all involve successive partitioning of cases into smaller and smaller bins based on one or more independent variables. A stopping criterion for the partitioning might be when the bins have 10 cases or less. For instance, one branch might be: if income < 32564 then if education < 14.2 then job satisfaction = 88.9, where 88.9 is the mean for the cases in that bin. Cutting points for branching the tree are set to minimize classification error as reflected in the residual sum of squares. As algorithms may produce an over-complex tree attuned to noise in the data, the research may "prune" the tree to trade off some increase in error in order to obtain less complexity in the tree. SPSS supports regression trees in its *AnswerTree* product. See Fox (2000b: 47-58).

Problems of regression trees. Use of automated tree algorithms commonly results in overfitting of trees (too much complexity, such that the branching rules seem arbitrary and unrelated to any theory of causation among the variables). This can be compensated in part by developing the tree for one set of data, then cross-validating it on another. Regression trees can be difficult to interpret because small changes in cutting points can have large impacts on branching in the tree. Branching is also affected by data density and sparseness, with more branching and smaller bins in data regions where data points are dense. In general, regression trees are more useful where the purpose is creating decision rules than when the purpose is causal interpretation.

GLM nonparametric regression allows the logit of the dependent variable to be a nonlinear function of the logits of the independent variables. While GLM techniques like logistic regression are nonlinear in that they employ a transform (for logistic regression, the natural log of the odds of a dependent variable) which is nonlinear, in traditional form the result of that transform (the logit of the dependent variable) is a linear function of the terms on the right-hand side of the equation. GLM non-parametric regression relaxes the linearity assumption to allow nonlinear relations over and beyond those of the link function (logit) transformation. See Fox (2000b: 58-73).

- **What is Poisson regression?**
  Poisson regression, which is used in modeling counts and rare events, is discussed in the section on logit modeling.

Bibliography

A highly recommended introduction to this topic is Schroeder, Sjoquist, and Stephan (1986).

Methodology

- Achen, Christopher H. (1982). *Interpreting and using regression.* Series: Quantitative Applications in the Social Sciences, No. 29. Thousand Oaks, CA: Sage Publications. Introduction notable for its admonitions against over-reliance on R^2 and beta weights rather than unstandardized b coefficients and level importance in interpreting independent variables.

- Achen, Christopher H. (1991). A polychotomous linear probability model. Political Methodology Society. Berkeley, CA. Achen argues that using in regression an ordinal variable with fewer than 5 categories introduces overly biased results.

- Allison, Paul D. (1999). *Multiple regression.* Thousand Oaks, CA: Pine Forge Press. An excellent introductory text.

- Berk, Richard A. (2003). *Regression analysis: A constructive critique.* Thousand Oaks, CA: Sage Publications.

- Berry, William D. (1993). *Understanding Regression Assumptions.* Series: Quantitative Applications in the Social Sciences, No. 92. Thousand Oaks, CA: Sage Publications. Excellent job at exactly what its title says, in more depth than given here.

- Breen, Richard (1996). *Regression Models: Censored, Sample Selected, or Truncated Data*, by Richard Breen. Quantitative Applications in the Social Sciences Series, No. 111. Thousand Oaks, CA: Sage Publications. Breen recommends Tobit when data are censored or sample selected.

- Bryk, A., S. W. Raudenbush, M. Seltzer, and R. T., Congdon (1988), *An introduction to HLM.* Chicago: University of Chicago. Describes the leading program for contextual analysis in regression.

- Chambers, William V. (1986). Inferring causality from corresponding variances. *Perceptual and Motor Skills*, Vol. 63: 475-478. This article lays the basis for the method of corresponding regressions.

- Chambers, William V. (1991). Inferring formal causation from corresponding regressions. *The Journal of Mind and Behavior*, Vol. 12, No. 1 (Winter): 49-70. This article sets forth the method of corresponding regressions, based on simulations and both natural and social science examples.

- Cohen, J. and P. Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* (2nd ed.). Lawrence Erlbaum Assoc. A widely used text.

- Draper, N.R., I. Guttman, and L. Lapczak (1979). Actual rejection levels in a certain stepwise test. *Communications in Statistics.* Vol. A8, pp. 99-105. The authors show how the alpha significance level becomes inflated during stepwise regression, increasing dramatically the chance of Type I errors.

- Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications.* London: Chapman & Hall.

- Fox, John (1991). *Regression Diagnostics.* Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No. 79. Provides a thorough review of methods of testing the assumptions of regression models.

- Fox, John (2000a). *Nonparametric simple regression.* Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No.130. Covers local polynomial multiple regression in detail.

- Fox, John (2000b). *Multiple and generalized nonparametric regression.* Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No.131. Covers local polynomial multiple regression, additive regression models, projection-pursuit regression, regression trees, and GLM nonparametric regression.

- Fuller, Wayne A.*Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage Publications. Reviews maximum likelihood regression before going on to logit and tobit models for categorical dependents.

- Hardy, Melissa A. (1993). *Regression with dummy variables.* Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 93.

- Iverson, Gudmund R. (1991). *Contextual analysis.* Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 81. Describes contextual analysis in regression.

- Jaccard, James, Robert Turrisi, and Choi K. Wan (1990). *Interaction effects in multiple regression.* Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 72.

- Kahane, Leo H. (2001). *Regression basics.* Thousand Oaks, CA: Sage Publications.

- Menard, Scott (1995). *Applied logistic regression analysis.* Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106.

- Miles, Jeremy and Mark Shevlin (2001). *Applying regression and correlation.* Thousand Oaks, CA: Sage Publications. Introductory text built around model-building.

- Schroeder, Larry D., David L. Sjoquist, and Paula E. Stephan (1986). *Understanding regression analysis: An introductory guide.* Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 57.

- Tabachnick, Barbara G. and Linda S. Fidell (2001). *Using Multivariate Statistics, Fourth Edition.* Boston: Allyn and Bacon.

Examples of Use of Regression in Public Administration

- Clingermayer, James C. and Richard C. Feiock (1997). Leadership turnover, transacrtion costs, and external city service delivery. *Public Administration Review*, Vol. 57, No. 3 (May.June): 231-239.

- Thompson, R. W. and E. S. Warren (1997). The role of regression methods in the determination of Standard Spending Assessments *Environment and Planning, C: Government and Policy* (UK). Vol. 15, No. 1 (Feb.): 53-72. Analyzes use of regression in public administration context of national allocation to local governmental units.

Back
